An Ontology-Based MicroRNA Knowledge Sharing and Acquisition Framework

Jingshan Huang, Xingyu Lu
School of Computing
University of South Alabama
Mobile, AL, U.S.A.
e-mail:huang@usouthal.edu
xl1001@jaguar1.usouthal.edu

Jiangbo Dang
Siemens Corporate Research
Siemens Corporation
Princeton, NJ, U.S.A.
e-mail: jiangbo.dang@siemens.com

Dejing Dou
Computer and Information Science Department
University of Oregon
Eugene, OR, U.S.A.
e-mail: dou@cs.uoregon.edu

Judith A. Blake
The Jackson Laboratory
Bar Harbor, ME, U.S.A.
e-mail: judith.blake@jax.org

William T. Gerthoffer
Department of Biochemistry & Molecular Biology
University of South Alabama
Mobile, AL, U.S.A.
e-mail: wgerthoffer@usouthal.edu

Ming Tan
Mitchell Cancer Institute
University of South Alabama
Mobile, AL, U.S.A.
e-mail: mtan@usouthal.edu

***Abstract*—MicroRNAs (miRNAs) play important roles in various biological processes by regulating their target genes. Therefore, miRNAs are closely associated with development, diagnosis, and prognosis for many diseases. The prediction of miRNA targets remains a challenging task for biologists because it involves an extremely large amount of data sources to be explored: to manually integrate information of identified targets and related information from various sources is time-consuming and error-prone; most of all, it is subject to biologists' limited prior knowledge. In this paper we investigated an ontology-based knowledge sharing framework to assist biologists in unraveling important roles of miRNAs in human disease in an automated and more efficient manner. (i) We developed the very first domain-specific ontologies in the miRNA field, Ontology for MIcroRNA Target (OMIT). (ii) According to the global metadata model defined in ontologies, heterogeneous data sources were annotated and seamlessly integrated and stored into a central Resource Description Framework (RDF) data repository. (iii) We then enabled ontology-based queries, instead of traditional SQL queries, by inferring new statements from RDF data triples. Consequently we were able to acquire hidden knowledge originally implicit and unclear, yet critical, to biologists.**

Keywords-semantic annotation and data integration; miRNA; ontology; knowledge acquisition; logic reasoning

I. INTRODUCTION

As a special class of RNAs, microRNAs (miRNAs) have been reported to perform important roles in a variety of biological processes by regulating target genes [16,20,22,18]. In particular, previous expression profiling has identified critical associations between miRNAs and human disease, such as cardiovascular disease, lung disease, and cancers. Unfortunately, the acquisition and prediction of target genes for miRNAs of interest remains a challenging task: substantial time and efforts have been spent in every search for available information in each small miRNA subarea. An example research scenario is presented as follows. Cancer patients' prognosis depends largely on their chemosensitivity [22]. Research has discovered that some specific genes increase the permeability of mitochondria membrane, which in turn leads to apoptosis. As a result, the patient's chemosensitivity will increase and the chemotherapy will be more effective. Certain miRNAs can regulate the aforementioned genes and thus affect cancer patients' prognosis. If biologists were able to identify such miRNAs, a breakthrough in treating drug-resistant cancers would result.

However, this identification is extremely difficult: not only do biomedical scientists need to manually extract hundreds of candidate targets from existing databases, but they also need to manually search and integrate related information for each candidate target, such as associated messenger RNA (mRNA) molecules, related protein functions, and affiliated pathways. The whole process is time-consuming, error-prone, and subject to biologists' limited prior knowledge because it involves an extremely large amount of data sources to be explored, including numerous miRNA target prediction databases, Gene Ontology (GO), PubMed/MEDLINE database, Gene Expression Omnibus (GEO) repository, and Kyoto Encyclopedia of Genes and Genomes (KEGG) PATHWAY database, etc. In addition, such a situation is further aggravated by great complexity and imprecise terminologies that characterize typical biological and biomedical research fields. A great deal of variety has been identified in the adoption of different biological terms, along with different relationships among all these terms. Such variety inhibits humans' effective knowledge acquisition. An example is *mir-143,* a member in the *mir-143~145* cluster, which is

well-known as a master regulator of the contractile phenotype of smooth muscle cells and as an important factor in early steps of differentiation of embryonic stem cells. Three target prediction databases, miRDB, PicTar, and TargetScan report 278, 216, and 404 genes respectively as targets for *mir-143*. It is very challenging, if not impossible, for biologists to manually search a total of such 898 candidate target genes (and from different sources), let alone to further search and integrate information from other sources on each gene. In fact, the situation could be even worse: biologists usually make use of additional target prediction databases in the miRNA area.

On the other hand, Semantic Web techniques, based on domain ontologies, have emerged as important tools in biomedical research. Ontologies are declarative knowledge models defining essential characteristics and relationships for specific domains. As a semantic foundation, ontologies render great help to biologists by formally defining domain knowledge. One of the most successful groups applying Semantic Web techniques into biomedical research is the GO Consortium [3] whose work has tremendously benefited protein structure and function studies. Besides GO, there exists a group of well-established bio-ontologies and related efforts, *e.g.,* the Unified Medical Language System (UMLS) [10], the Human Disease Ontology (HDO) [4], the Foundational Model of Anatomy (FMA) [21], Open Biological and Biomedical Ontologies (OBO) Foundry [14], and the National Center for Biomedical Ontology (NCBO) BioPortal [13]. *Unfortunately, there are not yet dedicated, specific miRNA-oriented ontologies.* The lack of domain ontological representation prevents miRNA researchers from taking advantage of automated data integration and logic reasoning, both of which are rendered by Semantic Web techniques and which will significantly enhance biologists' knowledge acquisition. Therefore, we aim to develop an ontology-based knowledge sharing framework to assist biologists in unraveling critical roles of miRNAs in human disease in an automated and more efficient manner.

The rest of this paper is organized as follows. Section II briefly reviews related work. Our methodology is described in detail in Section III, and we report the system implementation in Section IV, along with some discussions. Finally we conclude with future work in Section V.

II. RELATED WORK

A. MiRNA Target Prediction

Two categories of approaches have been developed for identifying miRNAs: experimental (direct biochemical characterization) and computational ones. After miRNAs have been identified through computational approaches, the next step is to perform experimental validation. Because direct experimental methods for discovering miRNA targets are time-consuming and costly, numerous target prediction algorithms have been developed, e.g., miRWalk [2] and miTarget [7]. Most of these algorithms adopt machine-learning techniques to construct predictors directly from validated miRNA targets. They typically depend on a combination of specific base-pairing rules and conservational analysis to score possible 3'-UTR recognition sites. In addition, systems like miRGator [12] and miRò [9] combine prediction results from different algorithms.

B. Semantic Annotation and Data Integration

There are three classes of semantic annotation/tagging systems: manual, semi-automatic, and automatic. Manual tagging systems (e.g., SemaLink [26]) require users to tag documents with a controlled vocabulary, which is a time-consuming process and requires deep domain knowledge and expertise. Semi-automatic tagging systems analyze documents and offer ontological terms, from which annotators may choose. Automated tagging systems analyze documents and automatically tag them with ontological entities. Zemanta [27] and SemTag [1] are such systems.

C. Semantic Web Techniques in Biological Research

Semantic Web techniques have been widely applied to biological research. The most successful example is the GO Consortium project [3], a major bioinformatics initiative since 1998. Consisting of three components, i.e., biological processes, cellular components, and molecular functions, in a species-independent manner, GO provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data, as well as tools to access and process such data. The NCBO [13] aims to support biomedical researchers in their knowledge-intensive work, provide online tools and a Web portal enabling researchers to access, review, and integrate disparate ontological resources in all aspects of biomedical investigation and clinical practice. A major focus of their work involves the use of bio-ontologies to aid in the management and analysis of data and knowledge derived from complex experiments.

An exploratory computing framework was proposed in our previous study [5, 11, 24] on domain ontologies to facilitate knowledge discovery in miRNA target prediction. Our previous papers focus on the feasibility analysis of the framework along with the construction of a preliminary version of domain ontologies.

III. METHODOLOGY

A. Overview

Figure 1 demonstrates the overall structure of OMIT framework.

- First, we developed OMIT ontologies to formally define miRNA domain knowledge. Ontologies standardize the terminology and define the domain knowledge contained explicitly or inexplicitly in data. The ontology development was driven by domain knowledge, popular upper ontologies, and existing bio-ontologies.
- We then created a central Resource Description Framework (RDF) data repository through semantic data annotation and integration. The repository contains a union of information integrated from distributed data sources and serves as a unified and consistent data layer for further analyzing data at the semantic level.

- Finally we incorporated ontology-based reasoning to enable users to perform complex semantic search/query among originally heterogeneous data sources. As a result, biologists will be able to acquire hidden, implicit knowledge from large amounts of data in an automated and more efficient manner.
- Biologists send queries through User Query Interface; RDF Query Engine retrieves predicted targets and related information integrated from distributed data sources, such as associated mRNA molecules, related protein functions, and affiliated pathways; RDF triples are sent to Inference Engine to perform logic reasoning; newly obtained knowledge, which was originally implicit and unavailable to biologists, is provided by the framework and sent back to users; biologists design experiments to validate computationally predicted targets from the framework and provide feedback to further refine the system.

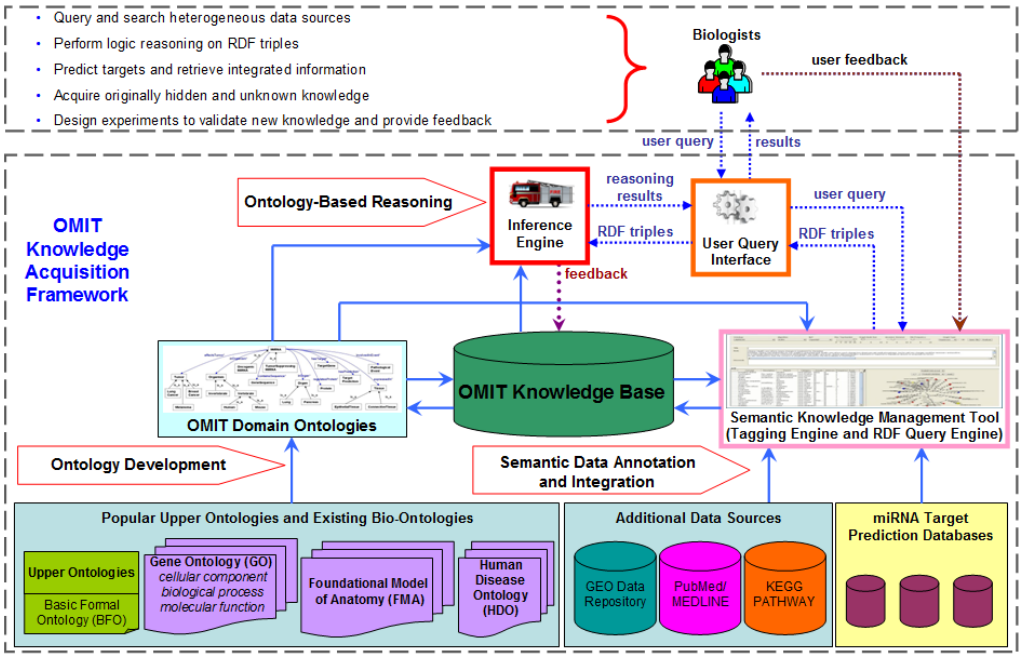

Figure 1. Overall structure of OMIT framework.

B. *Domain Ontology Development*

Many state-of-the-art miRNA target prediction systems have facilitated keyword search through syntax-based mechanisms supported by traditional relational databases. On the contrary, our framework places an emphasis on *data semantics* (intended meanings) rather than on *data syntax* (forms in which data are represented). The first step to handle well data semantics is to develop domain ontologies.

1) Ontology development principles: We have observed seven practices proposed by the OBO Foundry Initiative: (i) the ontology should be freely available; (ii) the ontology should be expressed using a standard language or syntax; (iii) successive versions of the ontology should be documented and tracked; (iv) the ontology should be orthogonal to existing ontologies; (v) the ontology should include natural language specifications of all concepts so they are readable by humans; (vi) the ontology should be developed collaboratively; and (vii) the ontology should be used by multiple researchers.

2) Knowledge-driven development procedure: Our ontology development approach was driven by domain knowledge and relied on various data sources: (i) popular upper ontologies (e.g., the Basic Formal Ontology); (ii) existing miRNA target prediction databases; and (iii) existing bio-ontologies (e.g., GO, HDO, and FMA). We have reused and extended a set of well-established concepts from existing bio-ontologies because we aim to (i) take advantage of the knowledge already contained in existing ontologies and (ii) reduce the possibility of redundant efforts. As illustrated in Figure 2, the whole development procedure is decomposed into five stages suggested in [25]: (i) specification of content, i.e., the range of concepts to be included in the ontology; (ii) informal documentation of concept definitions (typically carried out by domain experts); (iii) logic-based formalization of concepts and relationships between concepts (i.e., development of an "ontology" proper); (iv) implementation of the ontology in a computer language; and (v) evaluation of the ontology, including internal consistency and ability to answer logical queries. Besides, we emphasized that the ontology should be published in a form that is accessible to domain experts so that researchers can peruse and critique the contents of the ontology and consider what should be added or changed. All five stages in the ontology development are essentially ongoing and iterative because the process itself is complex. Besides, biologists' needs will change as their understanding of the domain evolves. In this iterative, knowledge-driven approach, both ontology engineers and

domain experts have been involved, working together to capture domain knowledge, develop a conceptualization, and implement the conceptual model. The ontology construction process has taken place over a number of iterations, involving a series of interviews, exchanges of documents, evaluation strategies, and refinements. We have adopted revision-control procedures to document the process for future reference. In addition, on a regular basis domain experts together with ontology engineers fine-tuned the conceptual model by an in-depth analysis of typical miRNAs, e.g., the *mir-143~145* cluster in the cardiovascular and respiratory systems and *mir-125b* in breast cancer.

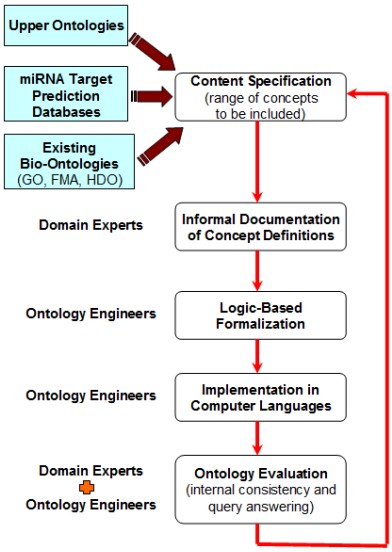

Figure 2. Knowledge-driven, iterative ontology development.

3) Ontology format and development tool: There are different formats/languages for describing ontologies, all of which are popular and based on different logics: Web Ontology Language (OWL) [17], Open Biological and Biomedical Ontologies (OBO) [14], Knowledge Interchange Format (KIF) [6], and Open Knowledge Base Connectivity (OKBC) [15]. We chose the OWL format that is recommended by W3C. OWL is designed for use by applications that need to process the content of information instead of just presenting information to humans. As a result, OWL facilitates greater machine interpretability of Web contents. We chose Protégé [19] as our development tool over other tools such as CmapTools and OntoEdit.

C. *Semantic Data Annotation and Integration*

According to the formal domain knowledge, including a global metadata model, defined in OMIT ontologies, heterogeneous data sources can be annotated and seamlessly integrated into a central RDF data repository. This data repository will serve as a unified and consistent data layer for further analyzing data at the semantic level.

1) Data sources: Data sources, containing structured, semi-structured, or unstructured data, to be integrated include: (i) a total of 17 miRNA target prediction databases (DIANA-microT, MicroInspector, miRanda, miRBase, miRDB, miRGator, miRGen, miTarget, NbmiRTar, PicTar, PITA, RNAhybrid, RNA22, TarBase, TargetScan, Vir-Mir, and ViTa); (ii) GEO repository; (iii) PubMed/MEDLINE database; and (iv) KEGG PATHWAY database. Because data are from heterogeneous sources the interoperability becomes an obstacle during the knowledge discovery. We adopted RDF, a standard model recommended by W3C for data interchange on the Web, to handle such a challenge. Due to its ability to facilitate data merging even if underlying schemas differ from each other, RDF allows structured, semi-structured, and unstructured data to be mixed, exposed, and shared across different applications, thus helping handle data interoperability.

2) Deep Annotation: Semantic data annotation is the process of tagging source files with metadata predefined in ontologies such as names, entities, attributes, definitions, and descriptions, etc. Herein, we use terms of "semantic annotation" and "semantic tagging" interchangeably. The annotation provides additional information contained in metadata to existing pieces of data. Metadata are usually from a set of ontological entities (including concepts and instances of concepts) predefined in ontologies. For unstructured data such as free text, we used a tagging engine to align them with ontological entities and generate semantic annotations. For structured data including database data, the annotation took two successive steps: (i) first we annotated data source schemas by aligning their metadata with ontological entities; (ii) according to annotated schemas we then transformed original data instances into RDF triples. We refer to such annotation as "deep" annotation – this term was coined by Goble, C. in the Semantic Web Workshop of WWW 02. It is necessary to annotate more than just data source schemas because there are situations where the opposite "shallow" annotation (i.e., annotation on schemas alone) cannot provide users with the desired knowledge. Taking the schema in the microRNA.org Web resource as an example: it combines a total of 172 fields into a single column. If users are only interested in, for example, knowledge pertaining to "AML-HL60" and "Astroblastoma-DD040800" instead of all 172 fields, it would then be extremely troublesome to retrieve the desired data for users if the conventional shallow annotation had been adopted. Following semantic data annotation, RDF triples can be indexed and accumulated into a central repository.

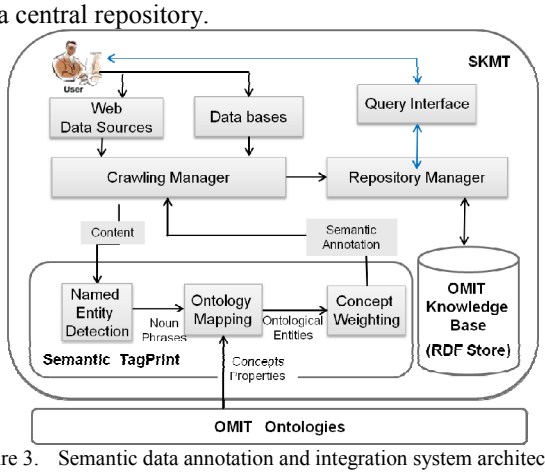

Figure 3. Semantic data annotation and integration system architecture.

3) System architecture: Figure 3 shows the system architecture of our semantic data annotation and integration subsystem. *Semantic TagPrint* is an automated semantic tagging engine that annotates free text using ontological entities. Three modules were designed for this component.

- *Named Entity Detection:* This module extracts named entities, noun phrases in general, from the input text. We adopted Stanford Parser [8] to detect and tokenize sentences, and assign Part-of-Speech (PoS) tags to tokens. Entity names were then extracted based on PoS tags.
- *Ontology Mapping:* This module maps extracted entity names to OMIT concepts and instances with two steps: *Phrase mapping* and *Sense mapping*. Phrase mapping matches the noun phrase of an entity name to a predefined concept or instance. Sense mapping utilizes a linear-time lexical chain algorithm to find the sense of the matched concept if it has several senses defined in ontologies. The lexical chaining algorithm disambiguates terms based on several ontological features such as *Hypernymy* and *Holonymy*.
- *Ontology Weighting:* This module utilizes statistical and ontological features of concepts to weigh semantic tags. Therefore, the input text will be annotated using the semantic with higher weights.

Crawling Manager collects original text and sends annotation results to *Repository Manager*, whose main role is to manage RDF repository (store) and to communicate with *Query Interface* (including *RDF Query Engine*). These components altogether provide a unified view over original data sources at the semantic level. Users will be guided by the *Query Interface* to automatically generate RDF queries across semantically integrated data sources. These queries will then be executed by a SPARQL-based RDF query engine. The semantics-based query improves the traditional keyword-based query in several ways. (i) Both synonymous terms (those having same meaning) and polysemous terms (those having different meanings) can be included to obtain more results that are relevant to the user query. (ii) Semantic relationships among terms often reveal extra clues hidden in disparate data sources. Such relationships can be explicitly discovered to further improve the quality of query answering. (iii) Given a set of candidate results to be returned to users, we can calculate the semantic similarity between each result and the user query using semantic features such as hypernym and holonym. We can then rank these results by their respective semantic similarities. Consequently, we are able to render users more accurate and more desired query results.

D. Ontology-Based Reasoning

The RDF data repository and the SPARQL-based query answering are not enough for the purpose of an effective and comprehensive knowledge acquisition. For example, the following two facts, "p53 must not be a direct target of *mir-885-5p*" and "*mir-21* upRegulates MalignantNeoplasm," do not exist in any one of the source miRNA target prediction databases, therefore, they will not be stored in the resultant RDF repository either. At the same time, such information is very helpful and thus preferable to biologists. In order to obtain the ability to acquire previously implicit knowledge and information, we incorporated an inference engine into the knowledge framework. Inference engines are also known as logic reasoners. Compared with traditional relational database techniques, inference engines that are specifically designed for OWL ontology models provide a more expressive method for querying, manipulating, and reasoning over available data sets. As a result, ontology-based queries, instead of traditional SQL queries, are thus made possible. Consequently, we will be able to acquire hidden knowledge and information that was originally implicit and unclear, yet critical, to biomedical researchers. With a logic reasoner, the OMIT RDF data repository works as a knowledge base.

We have preliminarily chosen Sesame [23] as our inference engine, which is an open-source Java framework for storing, querying, and reasoning (a.k.a. inferencing) with RDF and RDF Schema. Different types of reasoning tasks can be performed.

1) Subsumption reasoning: In formal logic, subsumption reasoning checks whether or not it is true that a concept (a.k.a. class) or a relationship (a.k.a. property) is subsumed by another concept or relationship. Due to the well-defined concept hierarchy specified in OMIT ontologies, we can readily retrieve subsumption relationships and integrate these relationships into query and search results before presenting them to users. Such additional conclusions will help biologists to generalize their findings to more model systems.

2) Contradiction reasoning: A logic reasoner can help check whether or not different components in the knowledge base (e.g., the concept hierarchy, rules, and instances) are consistent with each other. Note that incorrect or outdated information contained in original data sources is just one possible cause for contradicting or inconsistent situations; other events leading to contradiction and inconsistency include, but are not limited to, (i) human error; (ii) mistakes due to the annotation process; and (iii) errors happened during the data integration. Also note that contradiction or inconsistency may in fact reveal unique discoveries by a comparison between contradicting instances and original miRNA databases and studies.

IV. EVALUATION AND DISCUSSIONS

We implemented the OMIT framework in Java, and the only component that has not been completely included is SKMT. A project website (http://omit.cis.usouthal.edu/) was set up along with an online wiki to facilitate discussions among project partners and users. This dynamic site contains project-related materials (project goals, system design, instructional materials, case studies, application tools, and publications) via a Web portal, which serves as a gateway for further project dissemination. We expect the project website to create an immediate audience who can benefit from the early release of our prototype, and to provide valuable feedback to help us further refine the system.

A. *OMIT Domain Ontologies*

The most up-to-date OMIT ontologies contain a total of 397 concepts and 79 relationships. This version was submitted to the NCBO BioPortal and can be downloaded from http://bioportal.bioontology.org/ontologies/42873. Note that new terminology, such as *MiRNA*, *MiRNABinding*, and *ExperimentalValidation*, can be readily contributed to GO Consortium and other bio-ontology groups.

Figure 4 exhibits a screen shot from Protégé graphical user interface (GUI), with the concept *MiRNA* selected. Besides the schema information such as concept name, superClass/subClass relationships, and constraints on equivalentClass, some instances of *mir-143* binding sites are also shown in the figure.

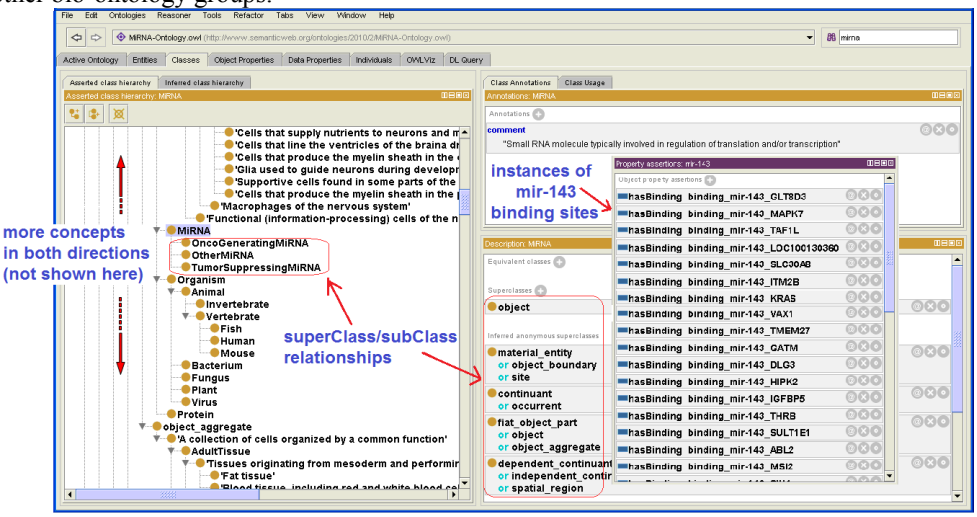

Figure 4. Protégé GUI screen shot that exhibits OMIT ontologies when the concept *MiRNA* is selected.

B. *OMIT Data Repository*

The current OMIT data repository contains a total of 3,146 facts (referred to as "axioms" in Protégé) integrated from information contained in miRanda, miRBase, miRDB, miRGator, miRGen, PicTar, PITA, RNAhybrid, TarBase, and TargetScan. These facts are specified in OWL, including 579 subclass axioms, 13 equivalent class axioms, 203 disjoint class axioms, 5 sub object property axioms, 2 inverse object property axioms, 29 object property domain axioms, 31 object property range axioms, 33 data property domain axioms, 32 data property range axioms, 393 class assertion axioms, 434 object property assertion axioms, 1011 data property assertion axioms, and 381 entity annotation axioms. In addition, we have converted all these facts into RDF triples and generate a central RDF data repository.

C. *User GUI and Query Answering*

A friendly user GUI was implemented in Java to answer biologists' query and search. We use the following example to demonstrate how the system works.

- The user specified the miRNA of interest along with all or a subset of properties of this miRNA (the top panel in Figure 5). Note that both selections were made through drop-down lists to minimize the user's input effort.
- The system automatically generated a SPARQL query statement in the back end corresponding to the user input; values for selected properties were then retrieved and populated in a separate panel (the bottom panel in Figure 5, which exhibits results when "*mir-143*" and five properties were chosen).
- The system displayed query results in a nested manner. For example, in order to view the details of the binding between "*mir-143*" and "VASH1" the user can click on the button located next to the binding list and a new panel (not shown here due to the space limit) will pop up with detailed information of the selected binding. The user was able to further acquire more details of "VASH1" – all integrated information regarding target "VASH1" was shown in another panel (Figure 6), where additional information was integrated, e.g., the target's *association terms* from GO. Note that the system was designed so that the user can easily go back and forth along links among ontological concept instances.

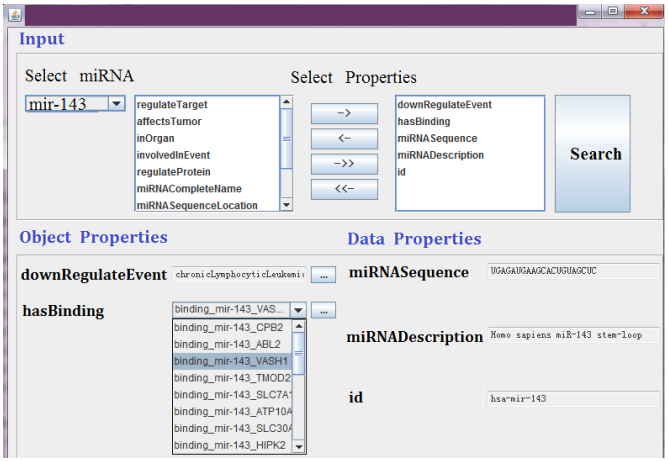

Figure 5. Friendly GUI for user query and search: Select miRNA and its properties and then obtain query results.

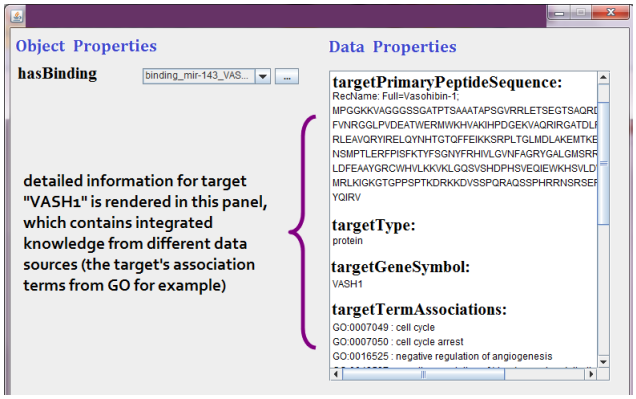

Figure 6. Additional information integrated in query results.

D. *Discussion*

The ultimate goal of this research is to enable biologists to acquire hidden, implicit knowledge from large amounts of data. The OMIT framework was designed upon domain ontologies, domain-specific knowledge and expertise can therefore be explicitly and formally encoded in the knowledge base. As a result, not only does the knowledge retrieved by the system contain a union of information integrated from originally heterogeneous data sources, but also inconsistent or conflicting facts from different sources will be identified. In addition, complex semantic search/query is made possible through logic reasoning.

1) Integrated knowledge: Query results retrieved by the OMIT system are regarded as integrated information in the sense that not a single data source alone involved in our framework contains such complete knowledge. For example, the list of instances of *mir-143* binding sites exhibited in Figure 5 was a result of combining predicted target bindings out of numerous target prediction databases. Moreover, additional information was obtained from sources other than miRNA target prediction databases: targets' *association terms* from GO ontologies, affiliated pathways from KEGG PATHWAY, etc.

2) Contradiction warning: Prior domain knowledge can be explicitly expressed as constraint rules in ontologies, "miRNAs frequently down regulate instead of up regulate their direct target genes" for example. During the incremental construction of the knowledge base, if instances violate predefined constraint rules, the inference engine will generate a warning message. Consequently, whenever this type of contradiction is going to take place, the reasoning mechanism is able to identify the scenario and prevent this contradiction from happening. Figure 7 exhibits a warning example when entering a new *upRegulateTarget* instance for *mir-21*. At this point, domain experts can make an arbitrary selection on either the new or existing instance.

3) Acquisition of hidden knowledge: Originally implicit and unclear knowledge can be critical to biologists during their knowledge acquisition, especially when they exploring a new research area. The notion of using ontologies as a foundation of our knowledge framework is for us to formally and explicitly encode the semantics of prior domain expertise into the knowledge base. On the contrary, it is very difficult, if not impossible, for traditional relational database techniques to handle the challenge of obtaining hidden knowledge because they focus on data syntax instead of semantics. For example, the following fact is contained in the knowledge base: "*mir-21* promotes hepatoCellularCarcinoma." At the same time, OMIT ontologies define "hepatoCellularCarcinoma" as an instance of the concept *Carcinoma*, which in turn is defined as a subclass of the concept *MalignantNeoplasm*. A new conclusion, "*mir-21* promotes MalignantNeoplasm," can thus be acquired by subsumption reasoning on the concept hierarchy contained in ontologies. Similarly, another conclusion, "*mir-21* promotes Tumor," can be readily obtained. These extra conclusions can help biologists to generalize their findings to more model systems.

Let us consider another scenario. Given a miRNA name, the OMIT system will search for its candidate targets from numerous target prediction databases, and the result is a set of targets for this specific miRNA. The system will further search in GO ontologies and find association terms for targets of interest. Note that association terms usually provide valuable information regarding biological experiments to be specifically designed for a particular target. For example, the term "ATP binding" often suggests that the target may involve in cellular metabolism or bioenergetics. Such prior knowledge can be formally encoded into ontologies via *partOf* or *involvedIn* relationships. Consequently, by reasoning on these relationships, the system can explicitly inform biologists of aforementioned valuable information, i.e., the suggestion that the target may involve in cellular metabolism or bioenergetics. In addition, if the system identifies other terms, "mitochondrion" for example, that are also related to the same biological process, i.e., cellular metabolism or bioenergetics in this example, the suggestion to biologists will be reinforced. A threshold can be set up so that when the reinforcement is above a certain value the system will automatically pop up a message to biologists.

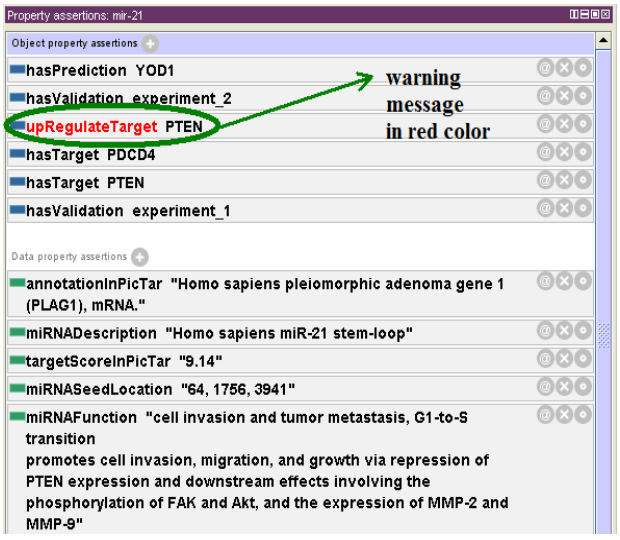

Figure 7. Contradiction warning by the inference engine.

Association terms can provide additional hints to biologists. Many terms can be categorized into one of the ten hallmarks of cancer, i.e., *EGFR*, *Cyclin-dependant kinase*, *Immune activating anti-CTLA4 mAb*, *Telomerase*, *Selective anti-inflammatory drugs*, *HGF/c-Met*, *VEGF signaling*, *PARP*, *Proapoptotic BH3 molecules*, and *Aerobic glycolysis*. Such categorization can be encoded into ontologies via *isa* or *partOf* relationships. Valuable information can then be obtained by reasoning on these relationships. The system can present biologists summarized cancer hallmark information for a particular miRNA of interest, i.e., the number of this miRNA's association terms in each and every hallmark category. Similar to the example in the previous paragraph, thresholds can be specified to measure the reinforcement degree inside each category. Moreover, detailed information for each hallmark category can be provided to biologists on demand, organized by either association terms (i.e., a list of targets under each term inside a category) or targets (i.e., a list of terms under each target inside a category).

V. CONCLUSION

Despite that miRNAs are closely associated with development, diagnosis, and prognosis for human disease, the miRNA target prediction remains a difficult task because to manually integrate information of identified targets and related information from various sources is time-consuming, error-prone, and subject to biologists' prior knowledge. We explored an ontology-based knowledge sharing framework to handle such a challenge. Our contribution can be summarized in three aspects: (i) we developed the very first domain-specific ontologies in miRNA field; (ii) new terminology can be readily contributed to GO Consortium and other bio-ontology groups; and (iii) ontology-based queries, instead of traditional SQL queries, were enabled by inferring new statements from RDF data triples through logic reasoning. As a result, the OMIT system provides biologists with enhanced knowledge acquisition in a more efficient manner, where not only is information integrated from various data sources, but also hidden knowledge originally implicit and unclear, yet critical, to biologists is presented. Consequently, our system will help biologists achieve faster discovery of miRNA functions.

This paper introduces the research motivation, methodology details, and system evaluation for the OMIT framework, respectively. An immediate future work is to integrate the SKMT component in the framework. We already had detailed design for this component. A second future research direction is to incorporate other reasoning tasks, *disjointWith* reasoning for example, so that more meaningful and valuable query-answering results may be obtained. Lastly, it will be worthwhile to research on the efficiency issue, especially when dealing with extraordinary large knowledge bases.

ACKNOWLEDGMENT

This research was supported in part by University of South Alabama Faculty Development Council (USAFDC) Grant (fund #: 144121).

REFERENCES

[1] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien, "Semtag and seeker: bootstrapping the semantic web via automated semantic annotation," Proc. 12th International Conference on World Wide Web (WWW 03), ACM, pp. 178–186, 2003.

[2] H. Dweep, C. Sticht, P. Pandey, and N. Gretz, "miRWalk-database: prediction of possible miRNA binding sites by "walking" the genes of 3 genomes," J. Biomed. Inform., 44(5):839-47, 2011.

[3] Gene Ontology, http://www.geneontology.org/index.shtml.

[4] HDO, http://bioportal.bioontology.org/ontologies/1009.

[5] J. Huang, C. Townsend, D. Dou, H. Liu, and M. Tan, "OMIT: A Domain-Specific Knowledge Base for MicroRNA Target Prediction," Pharmaceut. Res., Springer, August 2011.

[6] KIF, http://logic.stanford.edu/kif/.

[7] S. Kim, J. Nam, J. Rhee, W. Lee, and B. Zhang, "miTarget: microRNA target gene prediction using a support vector machine," BMC Bioinformatics, **7**:411, doi:10.1186/1471-2105-7-411, 2006.

[8] D. Klein and C.D. Manning, "Accurate Unlexicalized Parsing, " Proc. 41st Meeting of the Association for Computational Linguistics, 2003.

[9] A. Laganà, S. Forte, A. Giudice, M. Arena, P. Puglisi, R. Giugno, A. Pulvirenti, D. Shasha, and A. Ferro, "miRò: a miRNA knowledge base," Database: Journal of Biological Databases and Curation, 2009.

[10] D. Lindberg, B. Humphries, and A. McCray, "The unified medical language system," Meth. Inform. Med., 32(4):281-291, 1993.

[11] K. Murat, J. Dang, and S. Uskudarli, "Semantic TagPrint: Indexing Content at Semantic Level," Proc. 4th IEEE International Conference on Semantic Computing (ICSC 2010), Pittsburg, PA, USA, 2010.

[12] S. Nam, B. Kim, S. Shin, and S. Lee, "miRGator: an integrated system for functional annotation of microRNAs," Nucl. Acids Res., 36 (suppl 1): D159-D164, 2008.

[13] NCBO, http://www.bioontology.org/.

[14] OBO, http://www.obofoundry.org/.

[15] OKBC, http://www.ai.sri.com/~okbc/.

[16] P.H. Olsen and V. Ambros, "The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation," Dev. Biol., 216:671-680, 1999.

[17] OWL, http://www.w3.org/2004/OWL/.

[18] S. Pradervand, J. Weber, J. Thomas, M. Bueno, P. Wirapati, K. Lefort, G.P. Dotto, and K. Harshman, "Impact of normalization on miRNA microarray expression profiling," RNA, 15:493-501, 2009.

[19] Protégé, http://protege.stanford.edu/.

[20] B.J. Reinhart, F.J. Slack, M. Basson, A.E. Pasquinelli, J.C. Bettinger, A.E. Rougvie, H.R. Horvitz, and G. Ruvkun, "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans," Nature, 403:901-906, 2000.

[21] C. Rosse and J.L.V. Mejino, "The Foundational Model of Anatomy Ontology," Anatomy Ontologies for Bioinformatics: Principles and Practice, 6:59-117, London Springer, 2007.

[22] M. Selbach, B. Schwanhausser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky, "Widespread changes in protein synthesis induced by microRNAs," Nature, 455(7209):58-63, 2008.

[23] Sesame, http://www.openrdf.org/doc/sesame/api/org/openrdf/sesame.

[24] C. Townsend, J. Huang, D. Dou, S. Dalvi, P.J. Hayes, L. He, W. Lin, H. Liu, R. Rudnick, H. Shah, H. Sun, X. Wang, and M. Tan, "OMIT: Domain Ontology and Knowledge Acquisition in MicroRNA Target Prediction," Proc. 9th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2010), 2010.

[25] M. Uschold and M. Gruninger, "Ontologies: principles, methods, and applications," Knowl. Eng. Rev., 11(2):93-155, 1996.

[26] S. Wiesener, W. Kowarschick, and R. Bayer, "Semalink: An approach for semantic browsing through large distributed document spaces," Advances in Digital Libraries Conference, vol. 0, 1996.

[27] Zemanta, http://www.zemanta.com/.